Journal of Digital Imaging

Discerning Tumor Status from Unstructured MRI Reports—Completeness of Information in Existing Reports and Utility of Automated Natural Language Processing

Lionel T. E. Cheng,[1] Jiaping Zheng,[2] Guergana K. Savova,[2] and Bradley J. Erickson[1]

Information in electronic medical records is often in an unstructured free-text format. This format presents challenges for expedient data retrieval and may fail to convey important findings. Natural language processing (NLP) is an emerging technique for rapid and efficient clinical data retrieval. While proven in disease *detection*, the utility of NLP in discerning disease *progression* from free-text reports is untested. We aimed to (1) assess whether unstructured radiology reports contained sufficient information for tumor status classification; (2) develop an NLP-based data extraction tool to determine tumor status from unstructured reports; and (3) compare NLP and human tumor status classification outcomes. Consecutive follow-up brain tumor magnetic resonance imaging reports (2000–−2007) from a tertiary center were manually annotated using consensus guidelines on tumor status. Reports were randomized to NLP training (70%) or testing (30%) groups. The NLP tool utilized a support vector machines model with statistical and rule-based outcomes. Most reports had sufficient information for tumor status classification, although 0.8% did not describe status despite reference to prior examinations. Tumor size was unreported in 68.7% of documents, while 50.3% lacked data on change magnitude when there was detectable progression or regression. Using retrospective human classification as the gold standard, NLP achieved 80.6% sensitivity and 91.6% specificity for tumor status determination (mean positive predictive value, 82.4%; negative predictive value, 92.0%). In conclusion, most reports contained sufficient information for tumor status determination, though variable features were used to describe status. NLP demonstrated good accuracy for tumor status classification and may have novel application for automated disease status classification from electronic databases.

KEY WORDS: Natural language processing, unstructured, structured, radiology reports, tumor status

INTRODUCTION

The growing use of electronic medical records has resulted in vast stores of clinical information around the world that represent valuable resources for research and improving healthcare outcomes. However, the unstructured free-text format in which such electronic data are often stored poses significant challenges to expedient data retrieval. The inherent variability in content of unstructured reports may result in loss of information such as tumor progression. Further, even if pertinent information is present in the report, the complexities of human language render such reports less amenable to simple automated data retrieval.

Natural language processing (NLP) is an area dealing with computational methods for processing human language. It has been used as a main method of information extraction (IE), which aims to convert information residing in natural language into a structured format. Advances in both NLP

[1]From the Department of Radiology, Mayo Clinic, Rochester, MN, 55905, USA.

[2]From the Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, 55905, USA.

Correspondence to: Bradley J. Erickson, Department of Radiology, Mayo Clinic, Rochester, MN, 55905, USA; tel: +1-507-7748548; e-mail: bje@mayo.edu

Online publication 30 May 2009
doi: 10.1007/s10278-009-9215-7

and IE have allowed rapid data retrieval from electronic databases with accuracy comparable to human experts.[1,2] In the clinical domain, while NLP has proven utility[3] in detecting the *presence* of disease from unstructured reports,[4–15] it has not been evaluated as a tool for determining *progression* of disease.

The broad objective of this study was to determine if information regarding tumor progression could be accurately retrieved from unstructured follow-up magnetic resonance imaging (MRI) brain reports using NLP. Specifically, we first aimed to assess if the reports contained sufficient information for classification of tumor status. We next aimed to develop an NLP-based data extraction tool to detect changes in tumor status. Finally, we assessed if the NLP algorithm could retrieve information regarding tumor status from the unstructured reports with similar accuracy as an expert human interpreter.

MATERIALS AND METHODS

Ethics Approval

The study protocol was approved by Mayo Institutional Review Board.

Sample Selection

Consecutive MRI reports in the Mayo Clinic, Rochester, MN radiology report database from 1 Jan 2000 up to 1 Jan 2008 were screened for the following inclusion criteria:

1. Format: The report must be an unstructured free-text MRI brain examination report.
2. Indication: The MRI examination must be done for brain tumor evaluation. For our study, a "brain tumor" was defined as any of the following: brain tumor, brain cancer, glioma, meningioma, glioblastoma, astrocytoma, ependymoma, oligodendroglioma, brain lymphoma, brain metastases, and pituitary tumor.
3. Condition: The report must make reference to a suitable prior imaging study such as an earlier computed tomography or MRI brain examination.

MRI reports at our institution do not routinely have separate "observations/findings" and "impression/conclusion" sections. Instead, the reporting style is

Table 1. Study Consensus Guidelines for Manual Classification of Reports

Status indicator (regression, stable, progression)

This indicated the final overall tumor status compared to prior studies. Tumor size was the primary determinant of status, unless another feature was highlighted as indicating a status change despite stable tumor size. In the absence of specific reference to size, surrogate indicators (e.g. general statements on status, enhancement, signal intensity changes, mass effect, and presence of new lesions) were used to determine status.

Only changes from the *most recent* comparison study were considered. If a mix of 'stable' and another status ('progress' or 'regress') was present, then the net status change ('progress' or 'regress') was taken as the final status. If a mix of 'progress' and 'regress' statuses were present, the final status was classified as the worse (i.e. 'progress') status.

Magnitude indicator (mild, moderate, marked)

This indicated the qualitative extent of change, if any. The magnitude of change was classified as:

Mild if 'mild', 'slight', 'minimal', 'somewhat', 'small amount', 'subtle', 'appears to be some', 'tiny', 'partial', 'slow growth' or equivalent was used.

Moderate if 'moderate', 'modest', 'some' or equivalent was used. This was also the default classification if there was no specific mention of magnitude.

Marked if 'marked', 'significant', 'resolved', 'resolution', 'clearly', 'considerable', 'substantially', 'pronounced', 'large amount' or equivalent was used.

If several lesions with different change magnitudes were present, the greatest magnitude was chosen as the final magnitude.

Significance indicator (uncertain, possible, probable)

This indicated the subjective clinical significance of change, if any. The clinical significance was classified as:

Uncertain if 'uncertain', 'slight', 'subtle', 'unclear', 'not entirely typical of', 'indeterminate', 'non-specific findings' 'cannot be excluded' or equivalent was used. In the absence of specific significance indicators, a mild magnitude of change was tagged to an 'uncertain' significance.

Possible if 'possible', 'suggestive of', 'somewhat', 'benign rather than malignant', 'more consistent with post-therapy changes rather than neoplasm', 'could reflect', 'may represent', 'continued observation to assess', 'follow-up imaging to evaluate' or equivalent was used. This was also the default classification if there was no specific mention of significance and the magnitude of change was neither mild nor marked.

Probable if 'probably represents', 'worrisome for', 'concern that this represents', 'suspicious', 'concerning for', 'consistent with', 'compatible with', 'findings suggest', 'presumably indicating', 'findings indicate', 'findings likely reflect' or equivalent was used. In the absence of other specific significance indicators, a marked magnitude of change was tagged to a 'probable' significance.

If several lesions with similar status but differing significance were present, the greatest significance was chosen as the final significance.

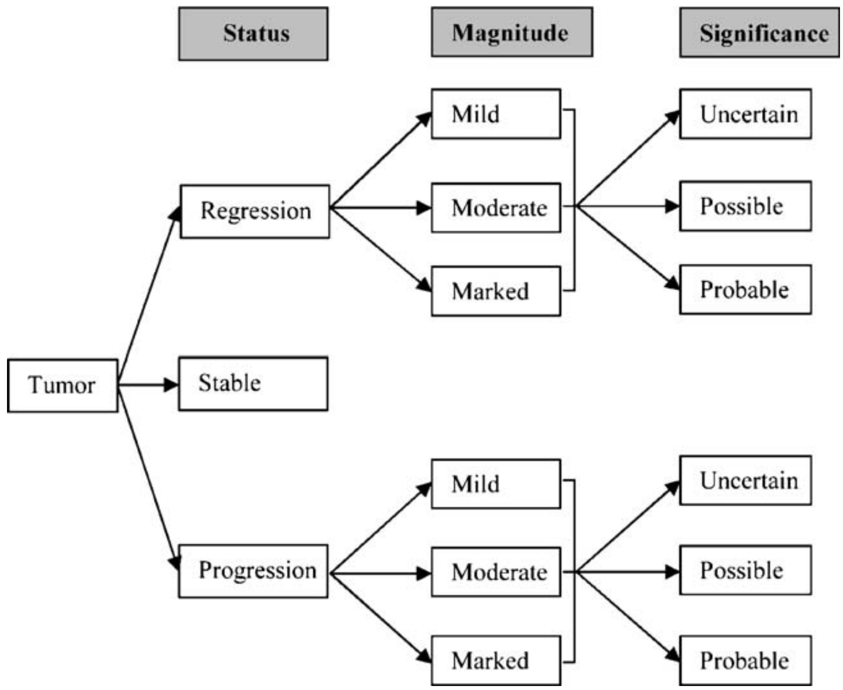


Fig 1. Classification scheme for radiology reports.

deliberately succinct, often consisting of key findings incorporated into an expanded impression.

Evaluating Information in Reports by Manual Classification

The selected reports were reviewed and annotated by a radiologist (LTC) according to consensus classification guidelines (Table 1) developed by the two authors (LTC and BJE) regarding disease status, magnitude of change, and the significance of change according to the classification scheme indicated in Figure 1. No additional clinical information, apart from data within the radiology report, was provided. Report annotation was performed using an open-source biomedical ontology editor (Protégé ver 3.2.1, Stanford Center for Biomedical Informatics Research, Stanford University School of Medicine, CA, USA) and a general-purpose text annotation plug-in tool (Knowtator ver 1.7.4, Center for Computational Pharmacology, University of Colorado Health Sciences Center, CO, USA). Ten percent of reports were randomly selected for blinded repeat annotation by the same radiologist 4 months after the initial annotation to evaluate intra-annotator agreement.

Developing an NLP-Based Data Extraction Tool

The reports were stratified by tumor status and divided into training (70%) and testing (30%) sets. Stop words (e.g., "if," "the," "by") had little lexical meaning and were removed. Content words were retained and underwent a stemming process using the Porter stemming algorithm[16] to reduce inflected variants to their stems (e.g., conversion of the word "reduction" to "reduce").

The NLP-based data extraction tool built for the task of discovering tumor status, magnitude of change, and significance of change combined statistical and rule-based methods (Fig. 2). The discovery of tumor status was cast as a classification task extending the support vector machines (SVMs) method,[17] while that for magnitude and significance was approached as a pattern-matching task. A simplified illustration of how an example report would be processed and analyzed by the NLP-based data extraction tool is given in Figure 3.

Discovering Tumor Status

A radiology report (document) could discuss multiple tumors and include additional information

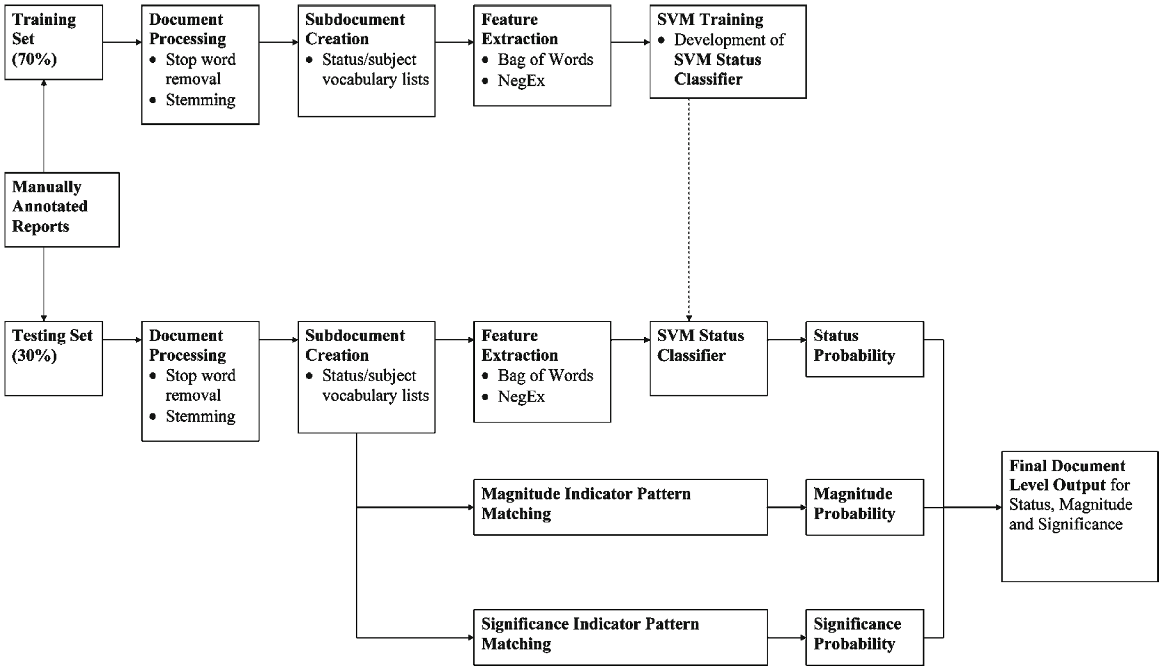


Fig 2. Development of NLP-based data extraction tool.

not directly related to tumor status. Therefore, the initial step in tumor status discovery was to identify narrative sections (hereafter referred to as *topic discourse units*) that contained information describing a single tumor.

Discourse Processing and Subdocument Creation The most important clue for the identification of topic discourse units was the description of status change. Based on the manual annotation outputs, vocabulary lists for tumor status (e.g., phrases indicating progression) and the tumor status subject (e.g., "mass," "abnormal signal") were compiled. Each document was then split into several subdocuments based on the occurrence of pairs of subject and status words from the two vocabulary lists. These pairs were restricted to within a maximum span of two adjacent sentences. Portions of the report with sentences not containing such pairs were evenly divided in terms of sentences and attached to the nearest subdocument. If a document described several tumors, each tumor description formed a separate topic discourse unit.

Feature Extraction Two types of features are extracted from each discourse unit. The first was the bag of words feature which allowed simplification of each document into a collection of words, disregarding grammar and word order. Each subdocument was represented as a bag of word stems in a vector space. The words in the bag were derived from sentences that had at least one tumor status or tumor status subject manually annotated. Thus, the vector was a series of binary values, with 1 for the presence of the word stem in the subdocument and 0 for absence.

The second feature, negation, was extracted using the NegEx algorithm.[18] Negation was common in the reports, and it was critical to distinguish between positive and negative mentions. For example, in the phrase "there was no significant growth," the "significant growth" is negated by the word "no." The NegEx algorithm focused on discovering anchor words and spanned a window on both sides of the anchor to detect negation markers. If a negation-stopping word occurred before the window was exhausted, then the scanning stopped there. The tumor status words were the anchors that fed into the NegEx algorithm with a window of six adjacent words. A tumor status was assigned a final value of negated if there was a negation word within the window and no intervening negation-stopping word was present.

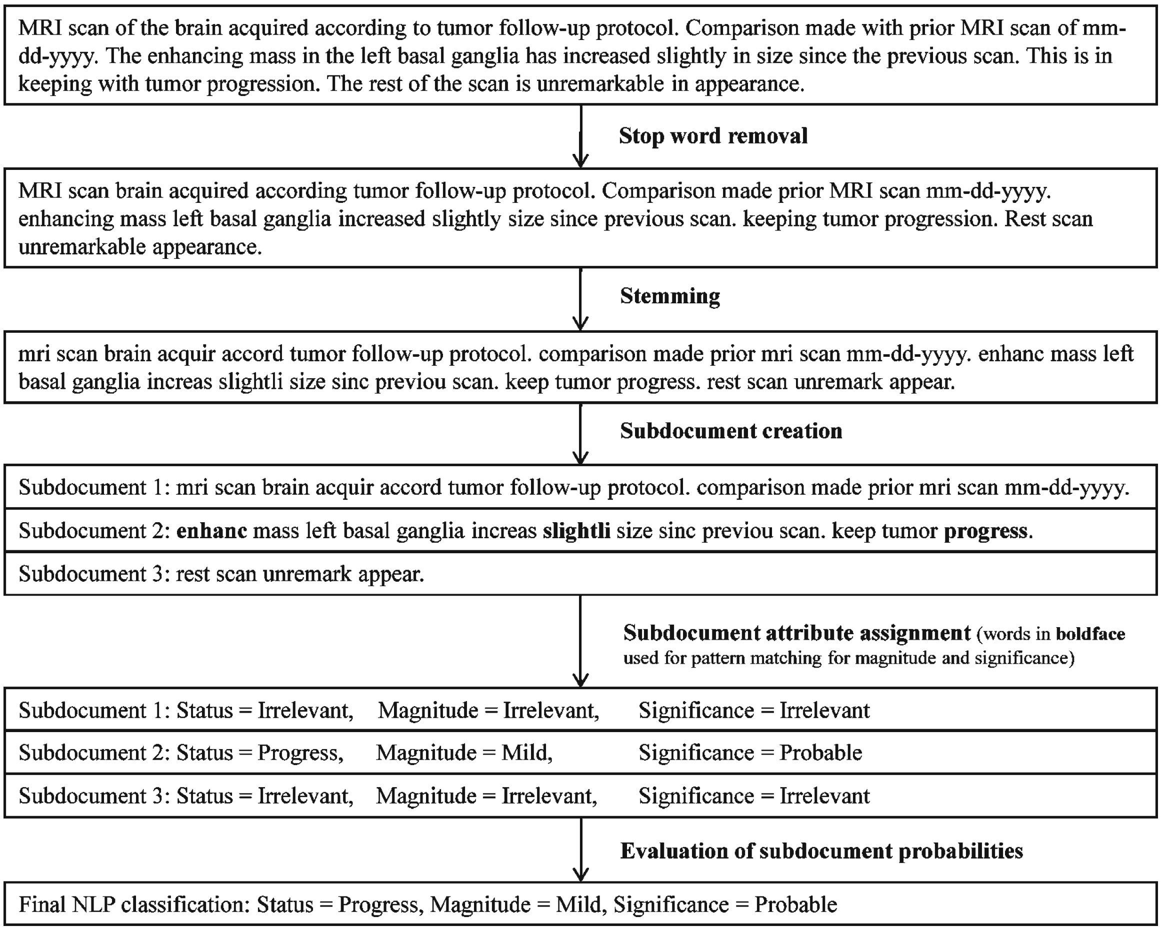


Fig 3. Simplified illustration of processing and analysis of an example report by the NLP-based data extraction tool.

SVM Training SVMs[17] are a machine learning technique (supervised learning method) for classification of data. Given training vectors ($\mathbf{x}_i$, y_i), i=1,…,n where $\mathbf{x}_i \in R^d$ and $y_i \in \{-1, 1\}$ as the class label, SVMs locate a hyperplane $\mathbf{w} \cdot \mathbf{x} - b = 0$ that maximizes the separation between the two classes. We used SVMs to build a classifier to discover tumor status. The feature vectors were constructed as described and a four-way SVM classifier with categories for *regression*, *stable*, *progression*, and *irrelevant* was trained. The LIBSVM[19] toolkit was used to extend SVM to support multi-category classification and enable probabilistic predictions.

Final Tumor Status Assignment Tumor statuses of subdocuments were assumed to be independent. The final document-level status probabilities were derived from both the subdocument-level probabilities (generated from the SVM toolkit) and the status assignment rules (Table 1). The probabilities were calculated as follows:

$$P(\text{irrelevant}) = \prod_{i \in \{\text{allsubdocs}\}} P(i = \text{irrelevant}) \tag{1}$$

$$\begin{aligned} P(\text{irrelevant} \cup \text{stable}) &= \prod_{i \in \{\text{allsubdocs}\}} P(i = \text{irrelevant} \cup \text{stable}) \\ &= \prod_{i \in \{\text{allsubdocs}\}} (P(i = \text{irrelevant}) + P(i = \text{stable})) \end{aligned} \tag{2}$$

$$P(\text{irrelevant} \cup \text{stable} \cup \text{regression}) = \prod_{i \in \{\text{allsubdocs}\}} (P(i = \text{irrelevant}) + P(i = \text{stable}) + P(i = \text{regression})). \tag{3}$$

Therefore,

$$P(\text{stable}) = P(\text{irrelevant} \cup \text{stable}) - P(\text{irrelevant}) \tag{4}$$

$$P(\text{regression}) = P(\text{irrelevant} \cup \text{stable} \cup \text{regression}) - P(\text{irrelevant} \cup \text{stable}) \tag{5}$$

$$P(\text{progression}) = 1 - P(\text{irrelevant} \cup \text{stable} \cup \text{regression}). \tag{6}$$

The final prediction at the document level was the *stable*, *regression*, and *progression* label with the highest probability.

Discovering magnitude and significance

Unlike tumor status descriptions, magnitude and significance had deterministic indicator patterns. This meant that apart from negation, there was little variation in the classification values that could be attributed to interaction between indicator patterns and other words in the same sentence. For example, the indicator pattern "compatible with" always indicated a *probable* value for significance, while "tumor cannot be excluded" always indicated an *uncertain* significance. Thus, these two classifiers were developed based on pattern matching, rather than the bag of words technique adopted for status classification. Each subdocument in a report was matched against a set of magnitude or significance indicators, while taking word order into account, and assigned a subdocument label according to the matching indicator. Subdocuments that lacked a magnitude indicator were assigned a default value of *moderate*. Subdocuments that did not have a significance indicator were assigned the same label as their immediate next subdocument with a significance indicator. If no significance indicator was present, an *irrelevant* significance was assigned.

For each possible configuration of subdocument tumor statuses in a report, the corresponding document-level magnitude/significance label was derived according to the classification guideline rules (Table 1), and the probability for this configuration was calculated. The document-level probabilities were derived by summing the same label assignments. For example, the probability of one report being assigned a final tumor status label of *mild* was computed as follows:

$$P(\text{mild}) = \sum_{i \in \{\text{all mild configs}\}} \prod_{j \in \{\text{all subdocs}\}} P(j). \tag{7}$$

The *magnitude* and *significance* probabilities were computed similarly. The final prediction was the one with the highest probability excluding *irrelevant*.

Comparing Human and NLP Classification Outcomes

The sensitivity, specificity, positive predictive value, and negative predictive value of the NLP classification outputs were calculated using the human expert classification as the gold standard. The statistical software used was JMP® 7.0 (SAS Institute Inc., Cary, NC, USA), which generated the main descriptive statistics, including kappa and Bowker values. Weighted kappa values were obtained using GraphPad QuickCalcs (GraphPad Software Inc., La Jolla, CA, USA). *F*-measures were calculated using Protégé (ver 3.2.1, Stanford Center for Biomedical Informatics Research, Stanford University School of Medicine, CA, USA).

RESULTS

A total of 778 MRI brain reports for 238 patients met the inclusion criteria, with character-

Table 2. Characteristics of Reports in the Manual Classification and NLP Development Groups

	Reports	Manual classification			NLP tool development		
Report characteristic	Overall ($N = 778$)	Classifiable ($N = 772$)	Unclassifiable ($N = 6$)	p value	Training group ($N = 541$)	Testing group ($N = 231$)	p value
Mean length (word count)	109	109	122	0.579	110	107	0.509
Incidental findings	49.4%	49.5%	33.3%	0.686	50.6%	46.8%	0.322
Spelling errors	11.6%	11.5%	16.7%	0.523	11.3%	12.1%	0.736
Fusion words[a]	14.8%	14.8%	16.7%	1.000	15.7%	12.6%	0.258

[a]Fusion words are combined words such as "inthe" or "lefthemisphere" that result from omission of a space between adjacent words

istics summarized in Table 2. The reports were prepared by 33 staff radiologists and had an average report length of 109 words (median 95; range 18–447). The number of reports per patient ranged from 1 to 22 (mean 3.3). Incidental findings were observed in almost half of the reports. These incidental findings were non-tumor-related observations such as sinusitis, vascular abnormalities, ischemic changes, normal variants, and incidental benign tumors unrelated to the neoplasm of interest.

Information in Unstructured Reports (Manual Classification)

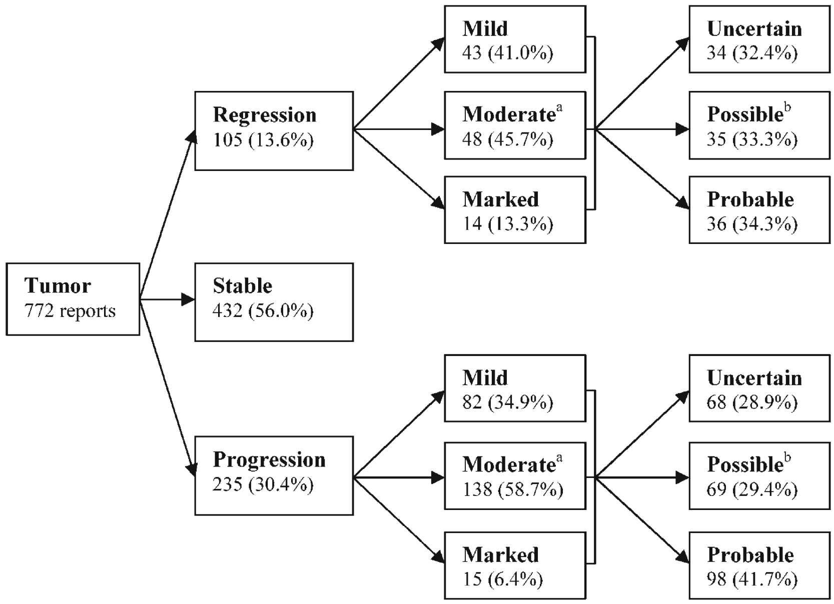
Out of 778 reports, six (0.8%) were unclassifiable despite having suitable comparison scans mentioned in the report (Table 2). Though these reports contained tumor descriptions, it was not possible to discern from the report text whether the findings constituted progression, regression, or no change, even after review by two radiologists (LTC and BJE). One unclassifiable document reported residual postoperative changes which hindered determination of tumor status. The unclassifiable reports had a greater mean report length compared to the classifiable reports, but this was not statistically significant. There was also no significant difference in the prevalence of incidental findings or spelling errors between the classifiable and unclassifiable reports.

In the 772 reports that were classifiable, tumor status was stable in 432 (56.0%), progressed in 235 (30.4%), and regressed in 105 (13.6%; Table 3). Reports could utilize either size, a surrogate indicator, or a combination of both types of indicators (size and surrogates) to describe tumor status. Surrogate indicators used were enhancement, signal change, new lesions, recurrent/residual tumor, or general statements on status. Overall, 557 (72.2%) reports utilized a surrogate indicator, while 242 (31.3%) used tumor size to describe status. This included 27 (3.5%) reports where both size and surrogate indicators

Table 3. Summary of Indicators Used for Determination of Tumor Status

Status indicator	Regression ($N = 105$)	Stable ($N = 432$)	Progression ($N = 235$)	Overall ($N = 772$)
Tumor size	55 (52.4%)	75 (17.4%)*	112 (47.7%)**	242 (31.3%)
Surrogate indicator[a]	50 (47.6%)	360 (83.3%)*	147 (62.6%)*,**	557 (72.2%)
Enhancement	37 (35.2%)	142 (32.9%)	95 (40.4%)	274 (35.5%)
T1 signal change[b]	4 (3.8%)	21 (4.9%)	13 (5.5%)	38 (4.9%)
T2 signal change	16 (15.2%)	85 (19.7%)	54 (23.0%)	155 (20.1%)
Mass effect	9 (8.6%)	8 (1.9%)*	12 (5.1%)**	29 (3.8%)
New lesion(s)[c]	4 (3.8%)	28 (6.5%)	53 (22.6%)*,**	85 (11.0%)
Recurrent neoplasm[c]	1 (1.0%)	127 (29.4%)*	4 (1.7%)**	132 (17.1%)
Residual neoplasm[c]	2 (1.9%)	132 (30.6%)*	1 (0.4%)**	135 (17.5%)
General statement[d]	5 (17.4%)	273 (63.2%)*	9 (3.8%)**	287 (37.1%)

*$p<0.05$ vs regression; **$p<0.05$ vs stable
[a]Features other than tumor size which depict tumor status
[b]Excludes signal change due to contrast medium enhancement
[c]Statements that indicate either the presence or absence of the features specified
[d]Overall statement about tumor status without reference to specific tumor features such as size, enhancement, signal change, etc.



[a]Reports without specific mention of magnitude were assigned a default 'moderate' magnitude.
[b]Reports without specific mention of significance were assigned a default 'possible' magnitude.

Fig 4. Outcomes of human annotation for classifiable reports.

were used. The types of status indicators used varied between different tumor status categories. In particular, the absence of recurrent/residual neoplasms and general statements about status were usually used to describe stability, while changes in status (regression or progression) tended to be reported as through descriptions of size and mass effect. As expected, the mention of new lesions was also associated with progression.

The magnitude of regression was classifiable in 60.0% (41.0% mild, 5.7% moderate, 13.3% marked) but unspecified in 40.0% (Fig. 4). Magnitude of progression was classifiable in 45.1% (34.9% mild, 3.8% moderate, 6.4% marked) but unspecified in 54.9%. Reports with status change contained variable degrees of significance (30.0% uncertain, 12.1% possible, 39.4% probable, and 18.5% unspecified). Reports without specific mention of magnitude or significance were subsequently assigned default values according to classification guidelines (Table 1).

Ten percent (77 reports) randomly selected for blinded repeat annotation by the same radiologist (LTC) gave high weighted kappa values for all categories (Table 4), indicating a high degree of intra-annotator agreement. There was no significant asymmetry for discordant classification outcomes (P>0.80 for all categories).

Comparison of NLP and Human Classification Outcomes

Compared to human classification for the test group (231 reports), NLP performed best for classification of tumor status, with an overall mean sensitivity and specificity of 80.6% and 91.6%, respectively (Fig. 5 and Table 5). Within the status subcategories, the highest NLP sensitivity was seen for classification of stability, while the highest specificity was obtained for classification of regression. The receiver operating characteristic

Table 4. Agreement Results

Classification category	Kappa	Weighted kappa	Bowker's test of symmetry
Intra-annotator agreement for human classification			
Status	0.98	0.96	$P=0.80$
Magnitude[a]	0.86	0.88	$P=0.80$
Significance[b]	0.82	0.87	$P=0.97$
Agreement between NLP and human classification			
Status	0.75	0.75	$P=0.58$
Magnitude[a]	0.68	0.71	$P=0.85$
Significance[b]	0.56	0.63	$P=0.41$

[a]Kappa values conditional on a correct status classification
[b]Kappa values conditional on a correct status and magnitude classification

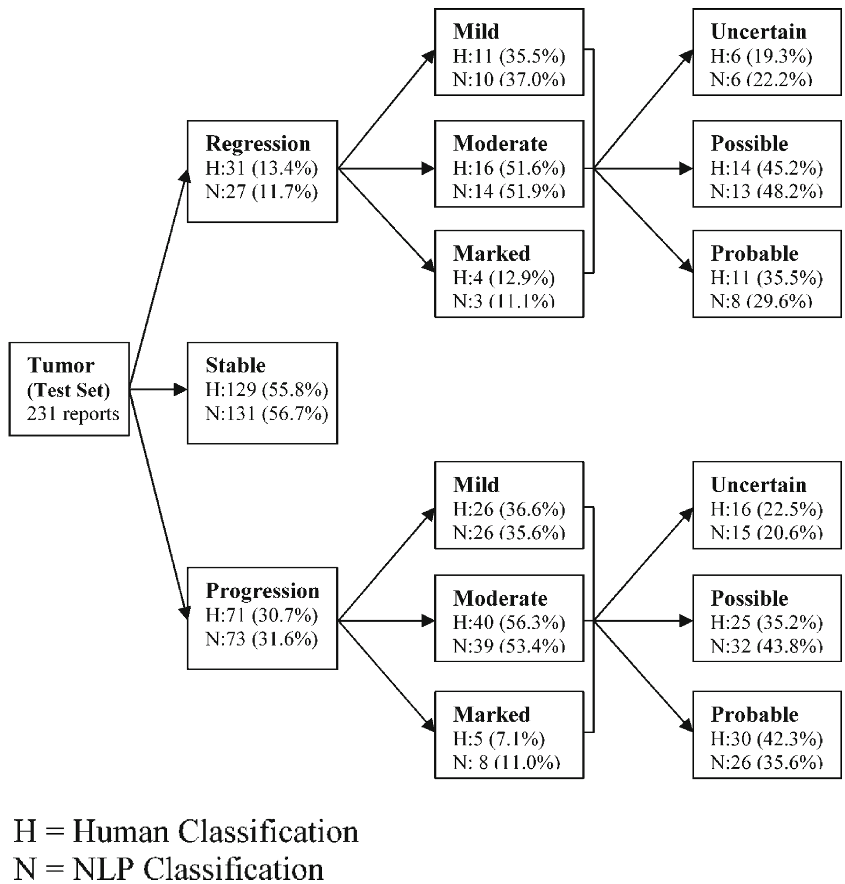


Fig 5. Comparison of NLP and human classification outcomes for reports in test set.

Table 5. Results of NLP Classification for All Categories

Category	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
Status				
Regress	64.5 (46.9–78.9)	96.5 (93.0–98.3)	74.1 (55.3–86.8)	94.6 (90.6–97.0)
Stable	89.9 (83.5–94.0)	85.3 (77.1–90.9)	88.6 (82.0–92.9)	87.0 (79.0–92.2)
Progress	87.3 (77.6–93.2)	93.1 (88.1–96.1)	84.9 (75.0–91.4)	94.3 (89.5–97.0)
Mean	80.6	91.6	82.4	92.0
Magnitude[a]				
Mild	85.7 (68.5–94.3)	90.7 (80.1–96.0)	82.8 (65.5–92.4)	92.5 (82.1–97.0)
Moderate	80.9 (67.5–89.6)	82.9 (67.3–91.9)	86.4 (73.3–93.6)	76.3 (60.8–87.0)
Marked	71.4 (35.9–91.8)	94.7 (87.1–97.9)	55.6 (44.4–73.3)	97.3 (90.5–99.2)
Mean	79.3	89.4	74.9	88.7
Significance[b]				
Uncertain	53.8 (29.1–76.8)	85.2 (73.4–92.3)	46.7 (24.8–69.9)	88.5 (77.0–94.6)
Possible	70.4 (51.5–84.1)	75.0 (59.8–85.8)	65.5 (47.3–80.1)	78.9 (63.7–88.9)
Probable	81.5 (63.3–91.8)	97.5 (87.1–99.6)	95.7 (79.0–99.2)	88.6 (76.0–95.0)
Mean	68.6	85.9	69.3	85.3

[a]Results conditional on correct status classification

[b]Results conditional on correct status and magnitude classification

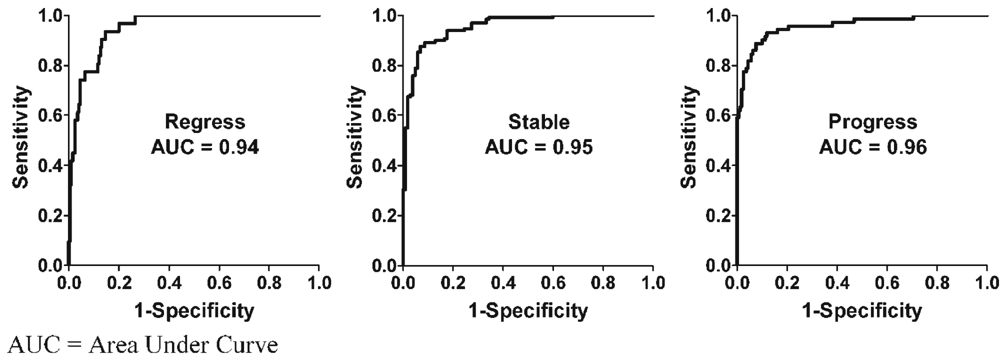


Fig 6. Receiver operating characteristic curves for tumor status determination by NLP.

(ROC) curves for NLP tumor status determination gave area under curve (AUC) values of at least 0.94 (Fig. 6). NLP performance metrics were lower for determination of magnitude and lowest for classification of significance. This trend was mirrored in the kappa values for agreement between NLP and human classification (Table 4). A similar pattern was observed for *F*-measures[20] of NLP compared to human classification. Macro *F*-measure scores of 0.81, 0.77, and 0.69 were obtained for status, magnitude, and significance respectively, while micro *F*-measure scores were 0.86, 0.82, and 0.72, respectively.

Subgroup analysis of NLP performance (Table 6) showed that for classification of tumor status, reports that were correctly classified were significantly shorter and more likely to contain general statements regarding tumor status. For magnitude of change, correctly classified reports had a significantly lower prevalence of surrogate status indicators (enhancement and statements about residual/recurrent neoplasms). For classification of significance, specific mention of tumor size was significantly associated with correct classification outcomes, while the use of surrogate status indicators (enhancement) had the opposite effect.

Table 6. Subgroup Analysis of Report Features Compared to NLP Classification Category Outcomes

	NLP classification outcomes								
	Status			Magnitude			Significance		
Report feature	+ (*N* = 198)	− (*N* = 33)	*p* value	+ (*N* = 67)	− (*N* = 48)	*p* value	+ (*N* = 48)	− (*N* = 67)	*p* value
Average report length	104	125	0.029*	127	130	0.794	125	131	0.599
Incidental findings	44.4%	60.6%	0.085	46.3%	52.1%	0.538	45.8%	50.8%	0.603
Spelling errors	11.1%	18.2%	0.249	10.5%	14.6%	0.504	6.3%	16.4%	0.148
Fusion words[a]	12.6%	12.1%	1.000	11.9%	12.5%	0.928	14.6%	10.5%	0.504
Tumor size	30.3%	30.3%	1.000	53.7%	35.4%	0.052	58.3%	37.3%	0.0257*
Surrogate indicator[b]	71.7%	72.7%	0.905	49.3%	70.8%	0.021*	45.8%	67.2%	0.0222*
Enhancement	32.8%	48.5%	0.081	28.4%	52.1%	0.0098*	25.0%	47.8%	0.0133*
T1 signal change[c]	4.6%	6.1%	0.660	4.5%	4.2%	1.000	6.3%	3.0%	0.648
T2 signal change	21.7%	27.3%	0.479	22.4%	22.9%	0.947	22.9%	22.4%	0.947
Mass effect	3.5%	6.1%	0.620	6.0%	4.2%	1.000	6.3%	4.5%	0.693
New lesion(s)[d]	13.1%	3.0%	0.141	20.9%	12.5%	0.242	20.8%	14.9%	0.410
Recurrent neoplasm[d]	15.7%	12.1%	0.794	0.0%	8.3%	0.0281*	0.0%	6.0%	0.139
Residual neoplasm[d]	17.7%	15.2%	1.000	0.0%	10.4%	0.0112*	0.0%	7.5%	0.074
General statement[e]	41.9%	9.1%	0.0002*	1.5%	8.3%	0.159	2.1%	6.0%	0.399

'+' and '−' indicate correct and wrong classification outcomes respectively. Statistically significant differences are marked by '*'

[a]Combined words that result from omission of a space between adjacent words

[b]Features other than tumor size which depicted tumor status

[c]Excludes enhancement after the administration contrast medium

[d]Statements that indicate either the presence or absence of the features specified

[e]Overall statement indicating tumor status without reference to a specific tumor feature such as size or enhancement

DISCUSSION

Our study demonstrates that the vast majority of unstructured radiology reports contain sufficient information to allow classification of tumor status by a human reader, although the linguistic indicators of tumor status varied significantly between reports. A novel NLP-based data extraction tool was developed and demonstrated utility in the classification of reports in terms of tumor status, change magnitude, and change significance. NLP classification outcomes had accuracy comparable to human expert classification, with the best performance seen for the classification of tumor status.

Completeness of Information in Unstructured Radiology Reports

Our findings support the many recognized limitations[21] of unstructured free-text radiology reports generated today. Although our study was limited to a specific clinical domain, the determination of brain tumor status, change magnitude, and change significance proved to be challenging even to a radiologist interpreter. Reports had typographical errors, different lengths, variable vocabularies, referred to multiple tumors, and included significant amounts of non-tumor information and incidental findings. We noted that variability in expression and interpretation was greater for the more subjective categories of magnitude and significance. For example, a lesion that is "slightly better appreciated" on a study could either refer to a real change (i.e., mild progression) or an apparent difference that was attributed to technical differences between studies (i.e., stability). The challenges of communicating doubt and certainty in radiological reports have been described perviously[22] and were borne out by the lower kappa values for the magnitude and significance categories. However, notwithstanding these challenges, there was a reasonably high level of reproducibility for human classification, which suggests that this categorization process could be successfully automated.

The case for structured reports and improved terminology in radiology has been made since the 1920s.[23–28] A structured report goes beyond standard headings (i.e., indication, findings, conclusion) to include consistent report content and a standard report language such as RadLex.[29] It is our view that using structured reports for tumor follow-up studies would result in more consistent and complete radiology reports. For example, though tumor size is advocated as a key measure of status,[30,31] the majority of reports in our study did not mention size. While the value of specific measurements is unclear for some tumors,[32] standardizing on which measurements are routinely included could help reduce confusion and improve communication. Compulsory data fields (including tumor size) customized according to the clinical question would help prevent the inadvertent exclusion of such important clinical information, reduce the ambiguity of terms used, and decrease variability in their interpretation. Furthermore, structured reports with standardized lexicons would be machine-readable, enabling decision support, semantic search, quality control, and rapid data mining to be more easily incorporated into the daily practice of radiology.

Utility of NLP for Tumor Status Determination from Unstructured Reports

Until the widespread use of structured radiology reports, NLP remains a promising option for rapid data retrieval from radiology report databases. A robust NLP classification tool could facilitate research by rapidly identifying specific patient or disease subgroups based on radiological findings. For example, all patients with changes in tumor status could be easily identified and studied for factors that contributed to the status change. In addition, "stable tumors" could be reviewed for changes too subtle for human detection. Such information can be used to improve automated decision support tools such as computer-assisted detection and diagnosis systems. NLP tools could also "screen" reports prior to finalization, prompting radiologists about important findings that may have been inadvertently left out. However, in view of the complexity and variability of language used in unstructured radiology reports, coupled with the existing error rate of our NLP tool, we feel that use should preferably be restricted to research purposes at the present time. The small but inherent misclassification rate could skew the characteristics of any subpopulation of patients retrieved by the NLP tool according to tumor status. Therefore, these limitations and causes of the errors should be further evaluated before use in a clinical scenario.

Our study showed that NLP was able to classify unstructured free-text neuroradiology reports according to tumor status with good accuracy. The NLP performance metrics were highest for tumor status, intermediate for magnitude, and lowest for significance. This was likely due to the increasing difficulty of determining magnitude and significance, which was also apparent during human classification. Though our NLP tool achieved high sensitivity for classification of stable tumors, we feel that further improvements are required before actual use in a research environment. It is our view that a usable screening tool should have both sensitivity and specificity exceeding 95%. Based on the ROC curve for classification of stable tumors, the current algorithm can only partially meet such criteria with either the combination of 95.4% sensitivity and 72.6% specificity or 79.1% sensitivity and 95.1% specificity.

Error analysis showed that a shorter report length and the presence of a general statement regarding tumor status were significantly associated with a correct NLP status classification. This may be due to a reduction of irrelevant information in short reports that could negatively influence the final classification outcome. A general statement on status was also more likely to be detected by the NLP algorithm. For magnitude and significance classification, the NLP tool performed poorer for reports that used surrogate markers of status other than size. This could be explained by the more variable vocabulary used to describe changes in surrogate features when compared to the more objective features used when reporting changes in size. It is noteworthy that spelling errors, fusion words, and incidental findings were not significantly associated with erroneous NLP classifications. This lends support to the utility of NLP for evaluating free-text medical reports, especially since it may not always be feasible to correct for such errors prior to NLP analysis.

Sensitivity of NLP classification was lowest for tumor regression, marked magnitude, and uncertain significance categories. Several factors may have contributed to this. Firstly, fewer reports were available for these categories, resulting in smaller training sets for the NLP algorithm. Secondly, it is possible that greater variability existed in how significance (including uncertainty) was expressed, as suggested by the lowest intra-annotator agreement levels obtained for classification of significance. Thirdly, uncertainty had the lowest priority amongst the significance categories in the classification scheme, meaning that a concurrent detection of uncertain and another significance value (including the default “possible” value) would always be resolved to the other value. Finally, when no explicit marker of significance was available, the discovery of uncertain significance was tagged to a mild magnitude. Therefore, any error in the discovery of mild magnitude would lead to downstream errors in the discovery of uncertain significance.

Several areas for potential improvement of the NLP algorithm were identified during the study.

For delineation of subdocuments in each report, our method relied on the existence of status–subject pairs identified using vocabulary lists with boundaries demarcated based on proximity. Because subject–status pairs were prevalent, each report tended to be split into many short fragments. Some of these subdocuments described change magnitude or change significance rather than tumor status. This resulted in the creation of irrelevant subdocuments that negatively influenced the subsequent probability contributions from relevant subdocuments. The lower performances of magnitude and significance classifiers could be partly attributed to this problem. Better identification of relevant subdocuments would help prevent the impact of irrelevant subdocuments on the final probability derivation and classification. Enhanced parsing, temporal reasoning,[33] and co-referencing are other areas where improvements in subdocument delineation could be achieved. These tools would enable NLP algorithms to better handle reports with multiple tumors, comparisons with more than one prior study, and lengthy reports with co-references that span multiple sentences.

Temporal reasoning remains a key challenge for automated medical data mining from medical reports.[33] The value of medical information is dependent on its temporal context. For example, tumor size in a report takes on greater significance when compared to a previously recorded size. This rate of change in size allows additional inferences regarding tumor aggressiveness or treatment efficacy to be made. However, automated temporal reasoning of unstructured radiology reports presents many challenges. For example, even the apparently simple definition of “time of study” can be difficult. The options include scan acquisition time, scan completion time, exam interpretation time, and report finalization time. Furthermore, there is no uniformly accepted format of representing day, month, and year in medical reports. Even if the format is standardized, time zone differences add an additional challenge, especially in the age of teleradiology where reports may be generated in a different time zone from where the examination was performed. Beyond definitions and representations, radiology reports may also make references to multiple prior studies without mentioning specific dates. This poses further challenges for automated temporal relationship discovery as a tumor may have both progressed and regressed when compared to different prior scans. The difficulty of automated temporal reasoning in radiology reports is also increased by temporal relations which rely on implicit event ordering with indicators such as "as compared to the scan taken last week" or "since admission" where no clear reference time point is given.

There are limitations for our study. Though SVMs have been successfully applied to text classification problems,[34] limitations exist for unbalanced datasets where class sizes differ significantly. Such unbalanced datasets are common in the medical domain where the important positive instance of a disease/outcome may be rare. In such cases, SVMs may favor the majority (negative) class and will still correctly classify most of the dataset even if the hyperplane is pushed towards the minority (positive) class. This is not preferred as false negative classifications of positive instances are less tolerable than false positive classifications. In our study, we addressed this limitation by conducting searches within a range of values to obtain empirically optimal parameters for the SVM hyperplane. Other methods have been suggested to reduce such errors.[35–37] Our findings are also limited to the subpopulation of patients with brain tumors with follow-up MRI studies. As imaging features and terminology vary between different tumor types and imaging modalities, our algorithm may not be applicable to different patient populations. The classification scheme used was formulated specifically for the study and had subjective components. While efforts were made to be in line with existing tumor status classification schemes,[30,31] the lack of uniformity across reports made complete alignment impossible.

CONCLUSION

Unstructured free-text radiology reports of tumor follow-up MRI brain examinations mostly contained sufficient information for determination of tumor status, though the features used to describe status varied significantly between reports. Almost 1% of reports studied could not be manually classified despite specific reference to prior exams in the report, over two thirds did not specify tumor size, and almost half did not report magnitude of change when either progression or regression was detected. We successfully developed an NLP-based data extraction tool using existing software that was able to determine tumor status from unstructured MRI brain reports with accuracy comparable to a human expert. Our findings show promise for novel application of NLP algorithms in ascertaining disease status from unstructured radiology reports.

ACKNOWLEDGMENTS

We would like to thank James Masanz and Karin Schuler for their valuable inputs and assistance with the annotation schema and Vinod Kaggal for the initial development of NLP algorithms.

REFERENCES

1. Automatic Content Extraction (ACE) Evaluation. Available at http://www.nist.gov/speech/tests/ace/. Accessed 17 Nov 2008

2. Message Understanding Conference (MUC) and Information Extraction. Available at http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html. Accessed 17 Nov 2008

3. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF: Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform, pp 128–144, 2008

4. Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD: Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med 122(9):681–688, 1995

5. Thomas BJ, Ouellette H, Halpern EF, Rosenthal DI: Automated computer-assisted categorization of radiology reports. AJR Am J Roentgenol 184(2):687–690, 2005

6. Elkins JS, Friedman C, Boden-Albala B, Sacco RL, Hripcsak G: Coding neuroradiology reports for the Northern Manhattan Stroke Study: a comparison of natural language processing and manual review. Comput Biomed Res 33(1):1–10, 2000

7. Melton GB, Hripcsak G: Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc 12(4):448–457, 2005

8. Dreyer KJ, Kalra MK, Maher MM, Hurier AM, Asfaw BA, Schultz T, Halpern EF, Thrall JH: Application of recently

developed computer algorithm for automatic classification of unstructured radiology reports: validation study. Radiology 234 (2):323–329, 2005

9. McCowan IA, Moore DC, Nguyen AN, Bowman RV, Clarke BE, Duhig EE, Fry MJ: Collection of cancer stage data by classifying free-text medical reports. J Am Med Inform Assoc 14(6):736–745, 2007

10. McCowan I, Moore D, Fry MJ: Classification of cancer stage from free-text histology reports. Conf Proc IEEE Eng Med Biol Soc 1:5153–5156, 2006

11. Pakhomov S, Weston SA, Jacobsen SJ, Chute CG, Meverden R, Roger VL: Electronic medical records for clinical research: application to the identification of heart failure. Am J Manag Care 13(6 Part 1):281–288, 2007

12. Savova GK, Ogren PV, Duffy PH, Buntrock JD, Chute CG: Mayo clinic NLP system for patient smoking status identification. J Am Med Inform Assoc 15(1):25–28, 2008

13. Denny JC, Miller RA, Waitman LR, Arrieta MA, Peterson JF: Identifying QT prolongation from ECG impressions using a general-purpose natural language processor. Int J Med Inform 78:S34–S42, 2009

14. Taira RK, Soderland SG, Jakobovits RM: Automatic structuring of radiology free-text reports. Radiographics 21 (1):237–245, 2001

15. Hripcsak G, Austin JH, Alderson PO, Friedman C: Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. Radiology 224(1):157–163, 2002

16. The Porter stemming algorithm. Available at http://tartarus.org/~martin/PorterStemmer/. Accessed 1 Nov 2008

17. Cortes C, Vapnik V: Support-vector networks. Mach Learn 20:273–297, 1995

18. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG: Evaluation of negation phrases in narrative clinical reports. Proc AMIA Symp, pp 105–109, 2001

19. LIBSVM—A library for support vector machines. Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/. Accessed 3 Nov 2008

20. Hripcsak G, Rothschild AS: Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc 12(3):296–298, 2005

21. Berlin L: Pitfalls of the vague radiology report. AJR Am J Roentgenol 174(6):1511–1518, 2000

22. Hobby JL, Tom BD, Todd C, Bearcroft PW, Dixon AK: Communication of doubt and certainty in radiological reports. Br J Radiol 73(873):999–1001, 2000

23. Hickey PM: Standardization of Roentgen–Ray Reports. Am J Roentgenol 9:442–445, 1922

24. Naik SS, Hanbidge A, Wilson SR: Radiology reports: examining radiologist and clinician preferences regarding style and content. AJR Am J Roentgenol 176(3):591–598, 2001

25. Dunnick NR, Langlotz CP: The radiology report of the future: a summary of the 2007 Intersociety Conference. J Am Coll Radiol 5(5):626–629, 2008

26. Reiner BI, Knight N, Siegel EL: Radiology reporting, past, present, and future: the radiologist's perspective. J Am Coll Radiol 4(5):313–319, 2007

27. Sistrom CL, Langlotz CP: A framework for improving radiology reporting. J Am Coll Radiol 2(2):159–167, 2005

28. Khorasani R, Bates DW, Teeger S, Rothschild JM, Adams DF, Seltzer SE: Is terminology used effectively to convey diagnostic certainty in radiology reports. Acad Radiol 10(6):685–688, 2003

29. Langlotz CP: RadLex: a new method for indexing online educational materials. Radiographics 26(6):1595–1597, 2006

30. Miller AB, Hoogstraten B, Staquet M, Winkler A: Reporting results of cancer treatment. Cancer 47(1):207–214, 1981

31. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, Verweij J, Van Glabbeke M, van Oosterom AT, Christian MC, Gwyther SG: New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. J Natl Cancer Inst 92(3):205–216, 2000

32. Galanis E, Buckner JC, Maurer MJ, Sykora R, Castillo R, Ballman KV, Erickson BJ: Validation of neuroradiologic response assessment in gliomas: measurement by RECIST, two-dimensional, computer-assisted tumor area, and computer-assisted tumor volume methods. Neuro Oncol 8(2):156–165, 2006

33. Zhou L, Hripcsak G: Temporal reasoning with medical data—a review with emphasis on medical natural language processing. J Biomed Inform 40(2):183–202, 2007

34. Joachims T: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: Lecture Notes in Computer Science. Berlin: Springer, 1998, pp 137–142

35. Veropoulos K, Campbell C, Cristianini N: Controlling the sensitivity of support vector machines. Proceedings of the International Joint Conference on Artificial Intelligence, pp 55–60, 1999

36. Lin Y, Lee Y, Wahba G: Support vector machines for classification in nonstandard situations. Mach Learn 46:191–202, 2002

37. Karakoulas G and Shawe-Taylor J: Optimizing classifiers for imbalanced training sets. In: Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 1999, pp 253–259